Large-scale integration of MicroRNA and gene expression data for identification of enriched microRNA-mRNA associations in biological systems

Chapter 20

Large-Scale Integration of MicroRNA and Gene Expression Data for Identification of Enriched MicroRNA–mRNA Associations in Biological Systems

Preethi H. Gunaratne, Chad J. Creighton, Michael Watson, and Jayantha B. Tennakoon

Abstract

The discovery of microRNAs (miRNAs) revealed a hidden layer of gene regulation that is able to integrate multiple genes into biologically meaningful networks. A number of computational prediction programs have been developed to identify putative miRNA targets. Collectively, the miRNAs that have been discovered so far have the potential to target over 60% of genes in our genome. A minimum of six consecutive nucleotides in the 5′-seed (nucleotides 2–8) in the miRNA must bind through complimentary base pairing to the 3′-untranslated (3′-UTRs) of target genes. Given the small sequence match required, a given miRNA has the potential to target hundreds of genes and a given mRNA can have 0–50 miRNA binding sites. The low-throughput nature of the query design (gene by gene or miRNA by miRNA) and a fairly high rate of false positives and negatives uncovered by the limited number of functional studies remain as the major limitations. Programs that integrate genome-wide gene and miRNA expression data determined by microarray and/or next-generation sequencing (NGS) technologies with the publicly available target prediction algorithms are extremely valuable on two fronts. First, they allow the investigator to fully capitalize on all the data generated to reveal new genes and pathways underlying the biological process under study. Second, these programs allow the investigator to lift a small network of genes they are currently following into a larger network through the integrative properties of miRNAs. In this chapter, we discuss the latest methodologies for determining genome-wide miRNA and gene expression changes and three programs (Sigterms, CORNA, and MMIA) that allow the investigator to generate short lists of enriched miRNA:target mRNA candidates for large-scale miRNA:target mRNA validation. These efforts are essential for determining false positive and negative rates of existing algorithms and refining our knowledge on the rules of miRNA–mRNA relationships.

1. Introduction

MicroRNAs (miRNAs) are small ~22 nucleotide noncoding RNAs that have been predicted to target >60% of the genes in our genome to mediate posttranscription gene silencing (1, 2). The

Silvia Monticelli (ed.), *MicroRNAs and the Immune System: Methods and Protocols*, Methods in Molecular Biology, vol. 667, DOI 10.1007/978-1-60761-811-9_20,

key determinants for miRNA–mRNA target associations lie in the 5′-seed region (nucleotides 2–8) in miRNA and the 3′-untranslated region (3′-UTR) of mRNA targets (2). The miRNA–mRNA target association is catalyzed mainly by the action of Argonaute (Ago) family of proteins in the RNA-induced silencing complex (RISC) (3). Base pairing of at least six consecutive nucleotides within the 5′-seed of the miRNA with the target site on the mRNA is reported to be required at a minimum. However, binding can occur through the entire length of the miRNA. miRNA–mRNA duplexes that form with perfect or near perfect complementarity have been shown to result in mRNA cleavage between nucleotides 10 and 11 (4) of the miRNA resulting ultimately in mRNA cleavage and decay (4, 5). By contrast, when binding occurs through imperfect complementarity, the mRNA target is generally kept intact and silencing occurs through translational repression (6).

With the advent of microarray and next-generation sequencing (NGS) technologies in the postgenome era, it is now possible to determine genome-wide miRNA–mRNA associations that are significant to specific cellular contexts or systems such as the immune system. A number of target prediction algorithms, which are primarily based on searches for matches between miRNA seed sequences and 3′-UTRs of genes, have been developed and freely available (7). Such programs offer users the possibility of quickly searching for potential targets on a miRNA by miRNA basis or potential miRNAs on a gene-by-gene basis. However, these approaches are too cumbersome and do not offer optimal solutions to integrate the glut of microarray (gene and miRNA expression) and sequencing (mRNA-seq and miRNA-seq) data that is becoming available on a daily basis. More recently, several groups have written programs and software packages to address this issue and offer solutions for the large-scale three-way integration of gene expression data, miRNA expression data and miRNA–mRNA target predictions (8). These programs offer the users the possibility of reaping the full benefit of these genome-wide studies. It is becoming increasingly clear that miRNAs are very different from the traditional transcriptional repressors that we are familiar with. Overexpression and loss-of-function studies suggest that most miRNAs have only a limited influence on their target genes (approximately two- to ten-fold repression) on its own. It appears that the main role of miRNAs is to fine-tune gene expression by coordinately downregulating multiple genes within and across pathways to integrate them into meaningful networks in relation to specific cellular states. The question then is what is the impact of global shifts in miRNA profiles on the transcriptome and proteome of a given cellular state. Furthermore, when aiming to assess the role of a given miRNA in relation to a specific biological process,

it is essential to consider its impact on all of its targets. Consequently, programs that integrate expression data with target prediction data are vital to understand the role of miRNAs in the immune system. In this chapter, we examine in detail three programs that allow the large-scale integration of genome-wide expression and miRNA target prediction data.

2. Materials

2.1. Gene Expression Data

2.1.1. Microarrays

While classical northern blotting and quantitative real-time PCR continue as techniques used for gene expression studies of single or small sets of genes over the past two decades, high-throughput microarray-based techniques have been increasingly applied in this field to measure several thousands of genes at a time. Microarray technology was first described by Schena et al. in 1995 (9). Over the years, DNA chip based technologies have widely demonstrated the power of this high-throughput parallel synthesis based method. Microarray DNA chips contain thousands of probes arranged on a regular pattern. Microarrays produce quantitative gene expression data based on relative dye intensities corresponding to DNA hybridized to probes immobilized on chips (10). A typical microarray-based experiment consists of preparing a DNA chip based on target DNAs, generating a hybridization solution containing a mixture of fluorescently labeled cDNAs, incubating fluorescently labeled cDNAs with DNA chip followed by data detection based on laser technologies, and finally computer assisted statistical testing and data analysis. To disseminate data analyzed by researchers for public use, microarray data can be stored in NCBI microarray data repository Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) (11) using a standardized framework, termed microarray markup language (MAML).

MAML employs a standard format to describe microarray experiment details, which include experimental design, array design, samples, hybridization procedures and parameters, images, quantitation, and controls.

Several commercial producers have introduced microarrays with different features. The microarrays available in the current market differ from one another in terms of the technologies utilized for fabrication and their probe design architecture. Some of the popularly known commercial manufacturers are given below:

Affymetrix GeneChip (http://www.affymetrix.com). Affymetrix was one of the first microarrays to appear in the market (12). Unlike in the case of traditional microarrays where cloning libraries are used for probe design, Affymetrix employs an in silico light directed synthesizing technology to produce probes on a glass chip (10).

Bypassing management of clone libraries and ability to synthesize highly ordered DNA oligomers in silico are the two distinct advantages of the Affymetrix design.

Agilent (http://www.agilent.com) employs an inkjet-based method to print whole cDNA or oligos on chips (13). Chips produced by Agilent offer 60-mer probes compared to the 25-mer probes offered by Affymetrix (10).

Nimblegen (http://www.nimblegen.com/) (14) uses a digital light processor to synthesize microarrays, apart from their use in transcriptome analysis. NimbleGen chips containing specific sequences are used to capture large genomic fragments, which can be subject to further analysis using Nimblegens GS FLX sequencing system (10).

CombiMatrix (http://www.combimatrix.com) offers custom arrays generated by a powerful computer-directed semiconductor microelectrode based on chip synthesis method, which can be programmed to generate a given array of oligonucleotides on chips (15). Signal detection can be carried out by either laser scanning or electrochemical methods (10).

Illumina bead array (http://www.illumina.com) (16) conventional microarrays are manufactured by spotting oligonucleotides on two-dimensional substrates (17). On the contrary, Illumina bead based arrays are produced by means of random assembly of bead pools on a patterned substrate (17). Illumina's technology offers higher oligo densities on their chips and thus higher throughputs by virtue of the intrinsic size of the beads and patterned substrates compared to conventional chips.

While array-based technologies and applications continue to grow, a plethora of information would be available for researchers through GEO in future. This would be a very valuable tool to facilitate cross-reference samples, identify signatures associated with disease, personalize medicine, and most importantly provide a global view of all biological processes through a platform for systematic in depth analysis of DNA and RNA variation.

2.2. miRNA Expression Data

2.2.1. MicroRNA Microarrays

The overall approach of miRNA profiling through microarrays remains similar to the approach employed in microarrays for gene expression profiling. Mature miRNAs are isolated and purified from tissue or cell samples using classical Trizol-based isolation or commercially available kits. The purified fragment of RNA is enriched and labeled. Array probes are designed by using locked nucleic acid (LNA) or chemically modified oligos and spotted on microarrays. Hybridization is then carried out and signal intensities measured using a laser scanner. Finally, quantification and data analysis is carried out using computer software. Unlike in the case of mRNA arrays designing arrays for miRNAs is challenging in that arrays must be designed to discriminate between the mature miRNAs and their precursors, miRNA microarrays should

be capable of detecting subtle differences of even a single-base difference of mature sequences (18). Short sequence length of 18–25 nt of mature sequences and wide range of melting temperatures (T_m) of mature miRNAs are significant problems in miRNA microarray design (19). In spite of the challenges, several miRNA microarrays have been designed and are currently available commercially. Synthetic oligonucleotides or cDNA fragments are used in miRNA microarray probe design. More recently, synthetic oligonucleotides with chemical modifications providing high molecular affinities facilitating hybridization have been employed. AT-rich probes are known to show lesser hybridization affinity compared to GC-rich probes (20). Higher degrees of sensitivity can be achieved by introduction of A/T analogs, which enhance overall duplex stability (21). Substitution of A and T with 2^0-O-methyl-2,6-diaminopurineand2^0-O-methyl-5-methyluridine, respectively, has shown two- to threefold increases in relative hybridization (22). LNAs first described by Wengel and coworkers in 1998 are a novel class of conformationally restricted oligonucleotide analogs, which show high thermal stabilities toward complimentary RNA and DNA (23). Chemically engineered LNAs have nucleotide analogs containing a bridging methylene group between C4′ and O2′ of the ribose ring (24). High thermal stabilities of LNAs bound to complimentary nucleic acid facilitates the design of short probes with excellent mismatch discrimination. Some of the commercially produced miRNA microarrays are discussed next.

2.2.1.1. Agilent miRNA Microarray (http://www.home.agilent.com)

Agilent miRNA microarrays are produced using unique chemically unmodified probes. Chemically unmodified oligos are immobilized on an array platform by means of a short stilt, and to the 5′ end of the anchored oligo a G residue is included and an extended hairpin attached, the 3′ end of the sample miRNAs are labeled by means of a Cy molecule attached to a C residue. When sample is introduced, hybridization takes place and the 5′ G residue of the probe complimentary to the 3′ Cy labeled C residue binds resulting fluorescence. The hairpin functions as a bridge connecting the 5′ end of the anchored oligo and the 3′ end of the hybridized miRNA. Agilent claims that the inclusion of the G residue to 5′ end of the probe increases stability of binding to target miRNA and the hairpin destabilizes probe hybridization to larger nontarget RNAs and hence provides a higher degree of specificity. Agilent's G44071A human miRNA microarray platform uses sequences from Sanger miRNA database (miRBase) version 12 and is capable of detecting unique 866 human and 89 viral miRNAs. Agilent also produces several arrays in the G44 series for human mouse and rat miRNAs, which use different versions of the Sanger database ranging from version 9.1 to 12.0 as the reference source for sequences. In Agilent miRNA

arrays, 40–60 mer unmodified oligonucleotides are directly synthesized on the array by Agilent's proprietary SurePrint inkjet technology. A unique feature of Agilent's technology is the use of end labeling instead of conventional polymerase based methods where sample nucleotide damage within the substrate has been an issue. End labeling is insensitive to nucleotide damage and is particularly advantageous when testing preserved or chemically treated samples. Agilent's platform requires only small input amounts in the 100 ng range of total RNA due to the high-yield end labeling method. As the labeling method does not require size fractionation or amplification, undesired bias introduced from these two steps is eliminated (25).

2.2.1.2. Exiqon LNA Microarrays (http://www.exiqon.com)

Exiqon uses melting temperature (T_m) matched LNA probes in their miRNA microarray design. Exiqon's miRNA microarrays are marketed under the name miRCURY LNA™. In addition to probes for miRBase sequences, which the Exiqon system uses as a reference for their microarrays, Exiqon arrays contain probes called mirPLUS™ capture probes, which target proprietary miRNAs that have been defined by Exiqon company through cloning and sequencing of human normal and diseased tissues. Through these proprietary sequence probes, scientists would be able to gain unique information about miRNAs, which have not been defined elsewhere. As of August 2009 in the Exiqon Web site, a typical miRCURY LNA™ was listed as being capable of capturing 854 mature human miRNAs, 80 mature viral miRNAs, and 428 mature Exiqon-defined human mirPLUS™ miRNAs (26).

2.2.1.3. Invitrogen (http://www.invitrogen.com)

Invitrogen offers the NCode™ Human miRNA Microarray Kit V3 and NCode™ Multi-Species miRNA Microarray Kit V2 as integrated miRNA profiling systems, which include reagents for RNA isolation labeling and array hybridization. As of the date of writing this chapter (30 August 2009), it was listed in the Invitrogen Web site that the Human miRNA Microarray Kit V3 contains probe sequences targeting nearly all of the known human miRNAs in the Sanger miRBase as well as probe sequences for 373 novel putative miRNAs. The Multi-Species version was listed as having probes for the Sanger miRBase Sequence Database, Release 9.0, for human, mouse, rat, *Drosophila melanogaster*, *Caenorhabditis elegans*, and Zebrafish. Each NCode™ microarray slide comes fully blocked and ready to use. In case where starting material has concentrations <500 ng total RNA or equivalent cells/tissue, Invitrogen provides a miRNA amplification kit called the NCode™ miRNA amplification system. Once the total RNA is extracted and ready for hybridization, labeling can be carried out using an NCode™. Rapid miRNA labeling system, which is

based on a poly-A tailing reaction on RNA molecules followed by ligation of dye labeled Alexa Flour™ DNA polymer by means of an oligoDT bridge. Invitrogen offers the choices of employing either preprinted or self-printed microarrays using NCode™ Human or multispecies microarray probes for experiments. Analysis of results can be performed by means of NCode™ profiler software. Invitrogen also provides a range of reagents under the NCode™ name for verification and further analysis of results by qPCR (27).

2.2.1.4. LC Sciences (http://www.lcsciences.com)

With the LC Sciences μParaFlo microfluidic miRNA Microarray chips assays can be performed with a minimum of 5 μg total RNA (28). The mirVana Isolation Kit (Ambion) is recommended. The small RNA (<300 nt) fraction is size fractionated with YM-100 Microcon Centrifugal Filter Device (Millipore) and 3′-extended with a poly-A tail by poly-A polymerase. An oligonucleotide tag is ligated to the poly-A tail for subsequent fluorescent dye staining. This platform allows dual labeling using Cy5 and Cy3 tags to label two RNA samples to be compared in dual-sample experiments. Hybridization is performed overnight on a μParaFlo microfluidic chip using a microcirculation pump (Atactic Technologies). On the microfluidic chip, each detection probe consists of a chemically modified nucleotide "coding" segment complementary to target miRNA (from miRBase, http://microrna.sanger.ac.uk/sequences/) or other RNA (control or customer defined sequences) and a spacer segment of polyethylene glycol to extend the "coding" segment away from the substrate surface. The detection probes are synthesized in situ with photogenerated reagent (PGR) chemistry on a Digital Light Projector (Texas Instruments) based synthesis system (29). Flexible DNA chip synthesis is gated by deprotection using solution photogenerated acids. The hybridization melting temperatures are balanced by adjusting length and chemical modifications of the detection probes (29). Hybridization is carried out in 100 μL 6× SSPE buffer (0.90 M NaCl, 60 mM Na_2HPO_4, 6 mM EDTA, pH 6.8) containing 25% formamide at 34°C followed by a stringent wash at 52°C. Hybridization images are collected with a laser scanner (GenePix 4000B, Molecular Device) and signal intensity values extracted using ArrayPro image processing software (MediaCybernetics). Data analysis is carried out by first subtracting the background and then normalizing with a cyclic LOWESS filter (locally weighted regression). For two-color experiments, the ratio of two sets of detected signals ($\log_2$ transformed and balanced) and p-values of the t-test are calculated; a p-value of less than 0.01 is used to select significantly differentially detected signal. Data classification is accomplished

by hierarchical clustering based on average linkage and Euclidean distance metric, and visualized with TIGR's Multiple Experimental Viewer (MeV) (30).

Quantile normalization on the channel values is used to normalize two-color data within each chip to make single channel values within and between arrays more comparable and to improve the multiarray data analysis. The single channel normalized values are used in subsequent data analysis. Construction of a dendrogram on the single channel values, both before and after normalization, is recommended to examine the effect of normalization on the treatment differences.

2.2.2. mRNA-seq: Next-Generation Sequencing

Completion of the human reference genome by the international human genome sequencing consortium and US-based Celera genomics was a cornerstone of human scientific endeavor. This achievement clearly paved way for a new exciting era of scientific research. The human genome sequencing project commenced in the year 1990; by 2000 a draft version of the human genome was made available and a completed version was released in the year 2003. During the human genome sequencing project era, the two widely used technologies were the original enzymatic dideoxy sequencing method pioneered by Fred Sanger and colleagues (31) and the Maxam and Gilbert method, which was described during the same year (32). The chemical degradation based Maxam and Gilbert method was particularly used in cases that were not easily resolved by the popular Sanger technique (33). As the human genome project progressed, the need for fast automated sequencers became imminent and companies with commercial interests were quick to step in to make improvements to the Sanger-based technique. In spite of the advances made in the Sanger technique through introduction of automated capillary sequencers, particularly the sample preparation steps, which involved cloning of sequences into bacterial artificial chromosomes (BACs) or yeast artificial chromosomes (YACs) and artifacts related to sample preparation remained obstacles of making Sanger-based sequencing a completely automatable high-throughput method. In view of this fact, several companies came up with novel sequencing technologies, which had massively parallel high-throughput capabilities enabling genome-scale analysis in a relatively short period of time. These sequencing technologies are termed NGS technologies. As of today, three platforms, namely Roche Applied Science 454 platform, the Illumina platform, and Applied Biosystems ABI SOLiD system are widely used in research laboratories. More recently, the Helicos single-molecule sequencing device, HeliScope was released to the market. A brief description of the 454, Illumina and SOLiD systems are given in the following paragraphs.

2.2.2.1. The 454 GenomeSequencer FLX Instrument (Roche Applied Science) (http://www.454.com/)

The 454 FLX pyrosequencer, which was released in 2004, was the first to be introduced to the market as an NGS (34). In pyrosequencing, each time a nucleotide gets incorporated to the nucleotide chain through a polymerizing reaction, pyrophosphate is released, and the released pyrophosphate leads to a series of downstream events, which results in the production of firefly luciferase (35). In the 454 system, DNA fragments are ligated with special adapters. One of the adapters facilitates binding of the DNA molecule to a bead. Beads containing single DNA fragments are subject to emulsion PCR and followed by a denaturation step. Initial amplification of sample DNA is necessary to generate sufficient signal strength in the sequence by synthesis step, which is subsequently carried out on beads containing copies of a given fragment immobilized on an optical fiber chip. In the 454 setup, each bead with its amplified fragment is individually addressable by a CCD camera at the fluorescence detection stage. In the sequence by synthesis stage, polymerase enzyme, primers, and a given labeled nucleotide of known identity are provided to each bead at a time, and the resultant fluorescence due to the pyrosequencing reaction is measured via the optical fibers equipped to a smart camera. By introducing labeled nucleotides of a given kind at each subsequent cycle of the polymerizing reaction, the nucleotides being incorporated to the growing fragment in each cycle can be detected by fluorescence measurement, and the sequence of each fragment can be decoded and assembled using sophisticated computer software. The 454 system is capable of detecting sequences in the 400–500 bp range and generates around 100 MB of data in a single run. A newer improved version of the 454 FLX called Titanium would provide a data output of around 500 MB. High costs of operation and generally low reading accuracy in homopolar stretches have been cited as drawbacks of the 454 system (33).

2.2.2.2. The Illumina (Solexa) Genome Analyzer (http://www.illumina.com/)

The Solexa sequencers were first introduced to the market in the year 2006 (36) and Illumina acquired Solexa in the year 2007 (33). The Solexa system is based on sequencing by synthesis method, which uses a technology called “Reversible termination”. The basic workflow of the Illumina platform involves five main stages. The initial step involves randomly fragmenting DNA and ligating adaptors to random fragments. The second step involves attaching DNA to a special glass slide and is followed by a third step, where solid-phase bridge amplification is carried out using unlabelled nucleotides. The fourth step involves denaturing amplified double-stranded DNA on the slide, and finally the fifth step involves carrying out a PCR using labeled nucleotides and photographing.

Unlike in the case of the 454 instrument where a single variety of nucleotide is incorporated in each cycle of the fluorescence

generating polymerizing step, the Illumina instrument introduces all four labeled nucleotides to the polymerizing reaction at once. However, due to a chemical modification of the nucleotides, each time a nucleotide gets incorporated into the growing DNA chain termination of polymerization occurs. At this stage, a smart CCD camera photographs fluorescence signals resulting from nucleotides, which got incorporated to each individually addressable amplified cluster of DNA fragments, which are generated at the bridge amplification stage. Once photographing of all clusters is completed, termination is reversed and another set of nucleotides are introduced, and once incorporation takes place, the reaction is terminated and clusters are photographed. Eventually all photographic data are analyzed and the sequences are assembled using computer software. The sequence read length achieved by this technology is around 35 bp, and an advantage of this system is its ability to generate huge amounts of data in a single run. The Illumina GA2 sequencers released in 2008 had the ability to generate around 1.5 GB of data in a single read setup and around 3.0 GB of data using a paired run. The ability of the instrument to generate massive amounts of data having short sequence lengths has made this instrument particularly well suited for small RNA based research, which generally does not demand long sequence reads. With various modifications in sample preparation and the use of different reagents, the Illumina platform can be used in a versatile fashion for ChipSeq and Bisulfite sequencing experiments as well.

2.2.2.3. The Applied Biosystems ABI SOLiD System (http://www3.appliedbiosystems.com/)

In contrast to the polymerase reactions used in 454 and Illumina methods, the Applied Biosystems SOLiD technology uses a ligation-based reaction to incorporate fluorescent-labeled nucleotides in the sequencing step (37). However, the Solid system shares similarities with 454 and Illumina as it utilizes an adapter ligated library and emulsion PCR on magnetic beads at the sample preparation stages. The overall work flow of the solid system can be summarized as follows. Initially, an emulsion PCR step is carried out on adapter ligated DNA fragments anchored to magnetic beads to provide sufficient fluorescence intensities during the detection step. The magnetic beads containing the amplified fragments are then transferred to a flow cell slide where a ligation reaction is carried out. The ligation reaction uses a primer, which attaches to the 5 prime end of the adaptor that immobilizes DNA fragments on the magnetic bead. DNA ligase and specific 8 mers whose fourth and fifth bases are specifically encoded with attached fluorescent labels are introduced to the reaction. Fluorescent detection is followed after each extending ligation step. After ligation and detection, a regeneration step in which the 8 mers including the fluorescent labels are removed is carried out and a primer corresponding to a single base displacement (n–1) from

the 3′ end of the adapter attaching the DNA fragment is introduced, and the ligation cycle is followed while the two encoded bases are read. Similar cycles are carried out starting with primers, which correspond to *n*–2, *n*–3, *n*–4, and *n*–5. In each cycle, the encoded two bases are interrogated and data stored. Finally, when all rounds of ligation have been completed, a computer builds the sequence by decoding the stored data as two base pair calls. A distinct advantage of this system is the use of two base pair encoding. As a result of two base pair encoding, it is possible to discriminate between base calling errors, true polymorphisms and single base deletions of the sequence by alignment against a high quality reference. The sequence length in the solid systems is defined in between 25 and 35 by the user. A sequencing run in a SOLiD system can yield 2–4 GB of DNA sequence data (35).

Information regarding the HeliScope instrument is available at http://www.helicosbio.com/ (38). There are also several other companies, which are in the process of manufacturing single-molecule based powerful sequencers employing *state-of-the-art* technologies. The following links provide information regarding these systems, which are either in the developmental phase or are ready to step into the market: VisiGen Biotechnologies (http://visigenbio.com/) (39), Pacific Biosciences (http://www.pacificbiosciences.com/index.php) (40), Sequenom (http://www.sequenom.com) (41), Oxford Nanopore Technologies, UK (http://www.nanoporetech.com/) (42), BioNanomatrix (http://bionanomatrix.com/) (43), and Complete Genomics company (http://www.completegenomics.com/) (44).

2.2.2.4. Small RNA Sequencing

The small RNA fraction is prepared for Illumina sequencing by the ligation of 5′ and 3′ RNA adapters according to Illumina's small RNA protocol, which can be found in the link http://www.illumina.com/downloads/rnaDGESmallRNA_Datasheet.pdf (45). Illumina's small RNA adaptors are ligated to the 5′ and 3′ ends of size selected <30 nt RNA. Adapter-modified DNA fragments will be enriched by PCR and further gel purified prior to sequencing. Small RNA sequencing for each sample is then performed using the Illumina Genome Analyzer (GA-2) according to the manufacturer's small RNA protocol. Typically, this protocol results in over 5–10 million small RNA sequence reads per sample per lane.

2.2.2.5. Bioinformatics Platform for Analyzing Small RNA Sequence Reads

A number of high-throughput computational pipeline have been developed for analyzing small RNA sequence reads generated by NGS technologies including Illumina sequencing (46). Our pipeline is described in (46, 47). For each sample, all unique sequence reads with a minimum read count of 10 are aligned to a reference set of miRNAs. The reference set is adaptable and currently consists of the 678 human and 472 mouse mature miRNA

sequences found in the miRNA database (miRBase version 11.0) plus 227 miRNA predictions from Berezikov et al. (48). It has been observed that the flexibility of DICER processing of the precursor miRNA produces a variety of sequence fragments, which may be active (49). To account for this, we perform a local Smith–Waterman alignment of each unique sequence read against each of the mature miRNAs in the reference, allowing for a 3-base overhang on the 5′ end and a 6-base overhang on the 3′ end. The alignments are scored such that a matching or overhanging base counts as two points and mismatches as –1. Each unique sequence read, which achieves a per-base alignment score of 2 (i.e., a perfect match) is associated with each mature miRNA for which it achieved that score. The read counts of all redundantly aligning reads are equally apportioned to all mature miRNAs to which they align.

2.2.2.6. Identification of Novel MicroRNAs

Each specimen is expected to generate multiple sequences that are not sufficiently similar to any known human miRNA. For this purpose, a number of algorithms have been developed to evaluate the likelihood that the unique sequence that does not align with a known miRNA is a putative novel miRNA. Our novel miRNA discovery pipeline is described in Creighton et al. (8, 47). First, all small RNA sequences that do not align with known miRNA precursors are mapped to the reference genome sequence of the species the small RNA is derived from (i.e., human, mouse, etc.). Each exact sequence match is fetched along with 100 bases flanking either side. These ~220-bp sequences are then tested for miRNA-like hairpin structure. The ~220-bp putative precursor sequences are evaluated with the Vienna package (www.tbi.univie.ac.at/RNA/) (50). Each of the unique sequences that map to a larger hairpin structures is tested for the Ambros criteria, which states that "authentic" miRNA sequences must map to one arm of a single-loop hairpin with a minimum free energy less than –25 kcal/mol (51, 52). Hairpins with overly large or unbalanced loops and unique sequences that map to the loop of the hairpin are rejected. After folding the read plus flanking sequence, the sequence is trimmed down to include only the plausible precursor and then folded again to ensure that the precursor was not artificially stabilized by neighboring sequence. Sequences appropriately placed in miRNA-like hairpins are considered to be "putative mature miRNAs" (pmms). Strong conservation of the mature miRNA, significant (but possibly weaker) conservation of the hairpin arm opposite the mature miRNA, and little or no conservation of the hairpin loop are considered a positive sign. Poorly conserved sequences are also considered since not all known miRNAs are conserved. If both the mature miRNA sequence and the miRNA-star sequence are found among the sequences, this candidate is

considered a definitive confirmed novel miRNA. If there is a substantial difference in abundance, the more abundant form is defined to be the mature miRNA and the less abundant form the miRNA-star sequence.

2.3. miRNA Target Prediction Programs

miRNA:mRNA target predictions for a number of different species are now available in public Web sites. We recommend the PicTar algorithm (http://pictar.bio.nyu.edu) (53), which uses predictions from Krek et al. (54); TargetScan algorithm (http://www.targetscan.org) (55), which uses predictions from Lewis et al. (56); and the miRanda algorithm (http://www.microrna.org) (57). The Sigterms software currently uses all three algorithms, and CORNA software is adapted for miRanda predictions (58).

Currently available algorithms are diverse, both in approach and in performance and all have room for improvement (7). A comparative description of some of the better known algorithms and their features are given below.

2.3.1. TargetScan (http://www.targetscan.org/)

TargetScan provides target predictions for mammalian/vertebrates offering predictions with site conservation consideration as well as without site conservation consideration (55). In predicting targets, the algorithm takes into account parameters such as stringent seed pairing, site number and factors influencing site accessibility. In the mode where site conservation is taken into account, there is an option to rank by preferential conservation instead of site context (7).

2.3.2. PicTar (http://pictar.mdc-berlin.de/)

PicTar provides target predictions for a wider variety of clades including mammalian/vertebrate, fly, and worm (53). The factors taken into consideration in this algorithm are stringent seed pairing for at least one of the sites of the miRNA, site number, and overall pairing stability (59). PicTar takes into consideration site conservation for all cases and does not offer a feature where target predictions can be done without taking conservation into account (7).

2.3.3. miRanda (http://www.microrna.org)

The miRanda algorithm is capable of making miRNA target predictions for mammal/vertebrate, fly, worm as well as additional species (56). In its criteria for target prediction, the algorithm takes into account site number, pairing to most of the miRNA, and moderately stringent seed paring (60).

2.3.4. PITA (http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html)

The PITA algorithm is capable of predicting miRNA–mRNA targets for mammalian/vertebrate, fly, and worm clades with site conservation consideration as well as without site conservation consideration (61). In its model for target predictions, PITA uses predicted site accessibility and stability as well as moderately stringent base paring and the number of sites (62).

3. Methods

Methods and software for integrating gene expression results with miRNA expression can help to maximally assess the role of miRNAs as integrators of genes into biologically meaningful networks. This is based on the fact that a given miRNA typically has predicted target sites in the 3′-UTRs of hundreds of genes and a given mRNA has multiple binding sites for several different miRNAs. In addition, genes that belong to specific pathways or networks are coordinately regulated. For all these reasons, it is essential that miRNA–mRNA association analyses are dealt with in the context of genome-wide changes in transcripts. The ultimate aim is to determine predicted miRNA–mRNA pairs that are correlated in expression in the context of a specific experiment. These could be the genes which are significantly differentially expressed when comparing two different biological states or genes that remain correlated in a treatment time course.

Current insight suggests that miRNAs exert their biologic effects by posttranscriptionally targeting gene expression; it follows that low expression of a given miRNA in a given system should conceivably cause a concomitant reversal of expression patterns for in silico predicted gene targets. Given this, we could define a miRNA–mRNA *functional* pair as consisting of a miRNA being predicted to interact with a given mRNA, where the two are also anticorrelated with each other in terms of expression. Public gene targeting prediction databases usually provide Web interface, where the user can look up predicted miRNA–mRNA functional pairs for a specific miRNA or gene of interest. In cases where the number of genes of interest (e.g., a set of genes arising from an expression profiling experiment) is in the hundreds, a gene-by-gene approach to looking up miRNA–mRNA pairs becomes impractical. Below we describe public software tools designed to make the task of integrating lists of genes and miRNAs easier.

3.1. CORNA (http://corna.sf.net)

CORNA (63) is an open-source package for the free statistical software R (http://www.r-project.org) (64) and allows scientists to analyze gene lists in the context of miRNA target predictions. In particular, when a list of genes and a list of miRNA target predictions are given, CORNA will carry out enrichment analysis to determine whether the gene list is enriched for particular miRNA targets more than that can be expected by chance. For example, the input gene list can come from a significant gene list from a microarray experiment or a biological pathway. Further methods within CORNA exist to test for significant associations

between miRNAs, pathways, and gene ontology (GO) terms and to display quantitative data associated with miRNA targets. CORNA employs three complementary statistical methods for enrichments analysis of relationships within lists of genes: the HyperGeometric test, Fisher's exact test, and the χ^2-test. Central to the flow of information through CORNA is the gene list, from which the user may test for significant miRNA target associations. The user may also start with a miRNA, find genes that are targeted by that miRNA, and then test that gene list for enrichment of KEGG pathways or GO terms. The user may also plot quantitative data associated with the targets of a particular miRNA. CORNA exclusively uses R vectors and data frames and includes functions for reading miRNA target data directly from miRBase (65) and microRNA.org (60). There are also helper functions to read gene and GO term data using biomaRt (66); microarray data directly from GEO (67); and pathway data directly from KEGG (68). A comprehensive tutorial exists at http://corna.sf.net.

3.2. Sigterms (http://sigterms.sourceforge.net)

Like CORNA, the Sigterms package allows the user to obtain miRNA–mRNA relationships for an entire set of genes (69). While CORNA runs with R, Sigterms consists of a set of Excel macros. The user enters a set of selected genes into an Excel "Annotation" workbook, which represents the entire set of genes on the gene profiling platform. The Annotation workbook can contain miRNA target predictions from one of the three commonly used algorithms (TargetScan, PicTar, and miRanda), as well GO annotation or other pathway information. Annotation workbooks for a given gene array platform representing human or mouse genes can be found at http://sigterms.sourceforge.net.

The user-provided list of genes is first entered into a Microsoft Excel document. The software will then look up the genes in the Annotation workbook to retrieve all miRNA–mRNA pairs for the given algorithm. For each miRNA, Sigterms computes an enrichment statistic that determines if the set of genes that are differentially expressed in the context of an experiment have binding sites more than expected by chance for that particular miRNA. Sigterms outputs the entire set of miRNA–mRNA pairs into an Excel worksheet; the user can then filter this worksheet for the *miRNAs* of interest (e.g., those miRNAs that are anticorrelated in expression with the genes). For computing the one-sided Fisher's exact tests for enrichment of a set of targets for a particular miRNA within the set of genes, the reference gene set determined by the complete probe set on a given array is used. To account for multiple testing of miRNAs, Monte Carlo simulation testing is performed using a 100 randomly generated gene sets. For a given gene set and a given target

prediction database, the number of miRNAs having a nominal significant p-value ($p < 0.05$) for target enrichment is computed for each of the 100 random tests. To calculate FDR, the average number of miRNA associations less than or equal to the given nominal p-value for the 100 random tests is used. The ultimate goal is to identify predicted targets within the gene set, that are enriched or overrepresented, which could help to implicate roles for specific miRNAs and miRNA-regulated genes in the system under study.

3.3. MMIA (http://129.79.233.81/~MMIA/mmia_main.html)

MMIA (which stands for "MicroRNA and mRNA integrated analysis") is a Web-based application meant to provide a "one-stop" combined analysis of the miRNA/mRNA input data for various pathway-associated gene sets (70). The user inputs mRNA expression data as a tab-delimited text file along with either a miRNA expression data table or a list of top expressed miRNAs. Given the user-defined statistical cutoff values, MMIA defines the differentially expressed genes and miRNAs from the data. Using miRNA prediction algorithms (TargetScan, PITA, and PicTar), MMIA then matches the upregulated or overexpressed genes with the downregulated or underexpressed miRNAs, and vice versa. MMIA can also generate heat maps of the data and search mRNA–miRNA pairs for pathway-related gene set enrichment. The MMIA software offers a convenient way for users to upload and analyze their data, though less flexible in how the analysis is carried out, as compared to CORNA or Sigterms.

Programs such as CORNA, Sigterms, and MMIA provide investigators without substantial bioinformatics support means by which they could make optimal use of their gene and miRNA expression data. The aim is to generate a list of miRNA–mRNA associations that are significantly correlated in the experiment of interest. The goal is to provide short lists of miRNA–mRNA pairs to be validated by direct biochemical assays, which establish that the miRNA–mRNA pair occurs in a duplex and coimmunoprecipitates in Argonaute complexes (71) and functional assays that demonstrate that the 3′-UTR of the mRNA is responsive to the cognate miRNA in luciferase or GFP reporter systems (72).

Acknowledgments

PHG and JBT are supported by a 1 R01 HL095382-01 grant. The authors would like to thank Gayani Rajapakse, Ana Hernandez, and Rajib Ghosh at University of Houston, Department of Biology and Biochemistry for their assistance in preparing this manuscript.

References

1. Freidman, R.C., Farh, K.K., Burge, C.B., and Bartel, D. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**, 92–105.
2. Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanisms, and function. *Cell* **116**(2), 281–97.
3. Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. (2004) Human Argonaute 2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell* **15**, 185–97.
4. Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. (2001) Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO* **20**(23), 6877–88.
5. Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* **305**, 1434–7.
6. Olsen, P. H. and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev Biol* **216**, 671–80.
7. Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell* **136**(2), 215–33. Review.
8. Creighton, C.J., Reid, J.G., and Gunaratne, P.H. (2009) Expression profiling of microRNAs by deep sequencing. *Brief Bioinform* **10**(5), 490–97.
9. Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–70.
10. Pariset, L., Chillemi, G., Bongiorni, S., Spica, V.R., and Valentini, A. (2009) Microarrays and high-throughput transcriptomic analysis in species with incomplete availability of genomic sequences. *N Biotechnol* **25**(5), 272–9.
11. NCBI microarray data repository Gene Expression Omnibus (Geo), Accessed on August 26, 2009 at http://www.ncbi.nlm.nih.gov/geo/.
12. Affymetrix GeneChip, Accessed on August 27, 2009 at http://www.affymetrix.com.
13. Agilent chip, Accessed on August 31, 2007 at http://www.agilent.com.
14. Nimblegen, Accessed on August 28, 2009 at http://www.nimblegen.com/.
15. CombiMatrix, Accessed on August 28, 2009 at http://www.combimatrix.com.
16. Illumina bead array, Accessed on August 27, 2009 at http://www.illumina.com.
17. Fan, J.B., Oliphant, A., Shen, R. et al. (2003) Highly parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol* **68**, 69–78.
18. Shingara, J., Keiger, K., Shelton, J. et al. (2005) An optimized isolation and labeling platform for accurate microRNA expression profiling. *RNA* **11**, 1461–70.
19. Yin, J.Q., Zhao, R.C., and Morris, K.V. (2008) Profiling microRNA expression with microarrays. *Trends Biotechnol* **26**(2), 70–6.
20. Lockhart, D.J., Brown, E.L., Wong, G.G., Chee, M.S., and Gingeras, T.R. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675–80.
21. Prosnyak, M.I., Veselovskaya, S.I., Myasnikov, V.A. et al. (1994) Substitution of 2-aminoadenine and 5-methylcytosine for adenine and cytosine in hybridization probes increases the sensitivity of DNA fingerprinting. *Genomics* **21**, 490–94.
22. Rampal, J.B., ed. (2001) DNA arrays: methods and protocols. In *Methods in Molecular Biology*, Vol. 170. Humana Press, Totowa.
23. Kumar, P., Singh, S.K., Koshkin, A.A., Rajwanshi, V.K., Meldgaard, M., and Wengel, J. (1998) The first analogues of LNA (locked nucleic acids): phosphorothioate-LNA and 2′-thio-LNA. *Bioorg Med Chem Lett* **8**(16), 2219–22.
24. Arora, A., Kaur, H., Wenger, J., and Maiti, S. (2008) Effect of locked nucleic acid (LNA) modification on hybridization kinetics of DNA duplex. *Nucleic Acids Symp Ser* **52**, 417–18.
25. Agilent human, mouse, and rat miRNA microarrays product note. Agilent Technologies, Accessed on August 27, 2009 at http://www.home.agilent.com.
26. Exiqon LNA microarrays, Accessed on August 29, 2009 at http://www.exiqon.com.
27. Invitrogen, Accessed on August 28, 2009 at http://www.invitrogen.com.
28. LC Sciences, Accessed on August 30, 2009 at http://www.lcsciences.com.
29. Gao, X., Le Proust, E., Zhang, H. et al. (2001) Flexible DNA chip synthesis gated by deprotection using solution photogenerated acids. *Nucleic Acids Res* **29**, 4744–50.

30. The institute for genomic research, Accessed on August 29, 2004 at http://www.tigr.org/.
31. Sanger, F., NIcklen, S., and Coulson, S. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–67.
32. Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**, 560–4.
33. Ansorge, W.J. (2009) Next-generation DNA sequencing techniques. *N Biotechnol* **25**(4), 195–203. Review.
34. The 454 genome sequencer FLX instrument (Roche Applied Science), Accessed on August 28, 2009 at http://www.454.com/.
35. Mardis, E.R. (2008) Next-generation DNA sequencing. *Annu Rev Genomics Hum Genet* **8**, 387–402.
36. The Illumina (Solexa) Genome Analyzer, Accessed on August 30, 2009 at http://www.illumina.com/.
37. The Applied Biosystems ABI SOLiD system, Accessed on August 30, 2009 at http://www3.appliedbiosystems.com/.
38. Helicos Biosciences Corporation, Accessed on August 28, 2009 at http://www.helicosbio.com/.
39. Visigen Biotechnologies Inc, Accessed on August 30, 2009 at http://visigenbio.com/.
40. Pacific Biosciences, Accessed on August 28, 2009 at http://www.pacificbiosciences.com/index.php.
41. Sequenom Inc, Accessed on August 28, 2009 at http://www.sequenom.com.
42. Oxford Nanopore Technologies, UK, Accessed on August 28, 2009 at http://www.nanoporetech.com/.
43. BioNanomatrix, Accessed on August 28, 2009 at http://bionanomatrix.com/.
44. Complete Genomics company, Accessed on August 28, 2009 at http://www.completegenomics.com/.
45. Illumina Inc, Accessed on August 28, 2009 at http://www.illumina.com/downloads/rnaDGESmallRNA_Datasheet.pdf.
46. Reid, J.G., Nagaraja, A.K., Lynn, F.C. et al. (2008) Mouse let-7 miRNA populations exhibit RNA editing that is constrained in the 5′-seed/cleavage/anchor regions and stabilize predicted mmu-let-7a:mRNA duplexes. *Genome Res* **18**(10), 1571–81.
47. Creighton, C.J., Nagaraja, A.K., Hanash, S.M., Matzuk, M.M., and Gunaratne, P.H. (2008) A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions. *RNA* **14**(11), 2290–6.
48. Berezikov, E., Plasterk, R.H.A., and Cuppen, E. (2002) GENOTRACE: cDNA-based local GENOme assembly from TRACE archives. *Bioinformatics* **18**(10), 1396–97.
49. Zamore, P. and Du, T. (2005) microPrimer: the biogenesis and function of microRNA. *Development* **132**, 4645–52.
50. Vienna package, Accessed on August 29, 2009 at http://www.tbi.univie.ac.at/RNA/.
51. Ambros, V., Bartel, B., Bartel, D.P. et al. (2003) A uniform system for microRNA annotation. *RNA* **9**(3), 277–9.
52. Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**, 911–40.
53. PicTar algorithm, Accessed on August 27, 2009 at http://pictar.bio.nyu.edu.
54. Krek, A., Grun, D., Poy, M.N. et al. (2005) Combinatorial microRNA target predictions. *Nat Genet* **37**, 495–500.
55. TargetScan algorithm, Accessed on August 29, 2009 at http://www.targetscan.org.
56. Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**(1), 15–20.
57. miRanda algorithm, Accessed on August 28, 2009 at http://www.microrna.org.
58. John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. (2004) microRNA target detection. *PLoS Biol* **2**(11), 363.
59. Lall, S., Grun, D., Krek, A. et al. (2006). A genome-wide map of conserved microRNA targets in *C. elegans*. *Curr Biol* **16**, 460–71.
60. Betel, D., Wilson, M., Gabow, A., Marks, D.S., and Sander, C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res* **36**, 149–53.
61. The PITA algorithm, Accessed on August 30, 2009 at http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html.
62. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat Genet* **39**, 1278–84.

63. Wu, X. and Watson, M. (2009) CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics* **25**(6), 832–3.
64. The R project for statistical computing, Accessed on August, 27, 2009 at http://www.r-project.org.
65. Griffiths-Jones, S., Saini, H.K., Dongen, S.V., and Enright, A.J. (2006) miRBase: tools for micro RNA genomics. *Nucleic Acid Res* **36**, 154–8.
66. Durinck, S., Morean, Y., Kasprzyk, A. et al. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**(16), 3439–40.
67. Barrett, T., Suzek, T.C., Troup, D.B. et al. (2008) NCBI GEO: mining millions of expression profiles – database and tools. *Nucleic Acids Res* **33**, 562–6.
68. Kanehisa, M., Araki, M., Goto, S. et al. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, 480–4.
69. Sigterms, Accessed on August 28, 2009 at http://sigterms.sourceforge.net.
70. miRNA and mRNA Integrated Analysis (MMIA), Accessed on August 28, 2009 at http://129.79.233.81/~MMIA/mmia_main.html.
71. Karginov, F.V., Conaco, C., Xuan, Z. et al. (2007) A biochemical approach to identifying microRNA targets. *Proc Natl Acad Sci U S A* **104**(49), 19291–6.
72. Ebert, M.S., Neilson, J.R., and Sharp, P.A. (2007) MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nat Methods* **4**(9), 721–6.